概念情報に基づく前置詞句係り先の曖昧性の解消

呉　浩東 †　古郡　廷治 †

英語前置詞句の係り先の曖昧性は文の構造的曖昧性の典型例をなすものである．本論文は，選好規則と電子辞書から得られる様々な情報に基づき，前置詞句の係り先を決定する手法を提案する．最初に，係り先を決める上での概念情報の役割と，それを電子辞書から抽出する方法を述べる．次に，概念情報をはじめ統語情報，語彙情報に基づく前置詞句の係り先を決める選好規則について述べ，選好的曖昧性解消モデルを提案する．このモデルでは選好規則によって一意的に係り先が決まらなかった場合，補助的に確率を使い，コーパスから得られるデータから確率計算をすることにより係り先の決定を行っている．使用頻度の高い12の前置詞句を含む2877文について行った実験では，86.9%の正解率を得た．これは既存の手法に比べ，2%から5%よい結果となっている．

キーワード:　前置詞句係り先，構造的曖昧性，選好ルール，概念情報，電子辞書

Resolving Prepositional Phrase Attachment Ambiguity Using Conceptual Information

HAODONG WU † and TEIJI FURUGORI †

Prepositional phrase (PP) attachment is a major cause of structural ambiguity in English. This paper proposes a method to resolve PP attachment ambiguities that is based on local and global preferential rules. We first explain how conceptual information is used in PP attachment process. We then describe the way the information, drawn from a conceptual dictionary, is incorporated into the preference rules. When the attachment can not be decided by preference rules, we use probabilistic estimation to predict the right attachment. After putting the disambiguation process in an algorithm and tracing it with a few examples, we show a disambiguation experiment and compare its result with those of existing work: the success rate we attained is better than those of other methods by 2 to 5 percent.

KeyWords: *prepositional phrase attachment, structural ambiguity, preference rule, conceptual information, electronic dictionary*

1　はじめに

英語前置詞句 (Prepositional Phrase, PP) の係り先の曖昧性は文の構造的曖昧性の典型例をなすものである．その解消は自然言語処理における難題の一つとしてよく知られている．この問題の解決法には，大略，構文構造に基づく手法，知識に基づく手法，コーパスに基づく手法，シソーラスに基づく手法がある．

† 電気通信大学 情報工学科, Department of Computer Science, University of Electro-Communications

構文構造に基づく手法は, 文の構成素の結び付き関係を構文情報によって決めようとするものである. この手法の代表例に, Right Association (Kimball 1973) と Minimal Attachment (Frazier1978) がある. Right Association では, 文の構成素は右側に隣接する句と結び付く傾向があると考え, Minimal Attachment では, 構成素はより大きな構造に係る傾向があると見る. こうして, 前置詞句は Right Association では名詞句 (NP) に, Minimal Attachment では動詞句 (VP) に係る傾向を示す.

構文構造に基づく手法には係り先を決めるのが簡単で, 意味分析や特定の知識に依存しないという利点がある. しかし, 係り先決定の正解率は低く実用性も低い (Whittemore, Ferrara and Brunner 1990).

知識に基づく手法は, 世界知識や対話モデルを用いて曖昧性の解消を試みるものである (Dahlgren and McDowell 1986; Jensen and Binot 1987 など). この手法では, ドメインを限定し, その範囲での知識の利用が有効にできれば高い正解率が得られる. しかし, 現在の知識表現技術では知識の獲得が難しく, コスト面の問題もある.

コーパスに基づく手法は, コーパスから諸種の情報を抽出した上で係り先の確率を計算し曖昧性を解消するものである. 近年, 大規模のタグつきコーパスの開発が進み, コーパスに基づく言語研究が活発になっている. Hindle ら (1993) の提案した語彙選好 (lexical preference) モデルは, コーパスから自動的に抽出した動詞, 目的語となる名詞, それと前置詞の出現頻度により LA(lexical association) score を計算し, その値によって前置詞句が動詞か名詞のどちらに係るかを判断している.

コーパスに基づく手法は曖昧性の解消の有望な方法であることが認められている. しかし, この手法は希薄なデータ (sparse data) の問題を抱えている. また, 現状ではコーパスの資源や計算量の膨大さの問題もある.

シソーラスに基づく手法は, シソーラスや機械可読辞書の情報を利用し, あるいはシソーラスと例文を利用することによって, 前置詞句の曖昧性の解消を行うものである (Jenson and Binot 1987; Nagao 1992; 隅田ら 1994 など). この手法では特定のドメインで高い正解率を達成している. しかし, ドメインを限定しない場合, 単語の多義性によって係り先の決定率が著しく低下する傾向がある. また, シソーラスや辞書にはカバーする情報が分野によって不均一であることや意味の粒度の問題もある.

本稿では概念情報に基づく曖昧性の解消手法 (Conceptual Information Based Disambiguation, CIBD) を提案する. ここでは, まず言語知識と曖昧性解消に使っている世界知識から, いくつかの一般的な係り先決定ルール (選好ルールとよぶ) を抽出する. 選好ルールは係り先決定に際し, 概念情報をはじめ, 語彙情報, 構文情報と共起情報を利用している. もし, 選好ルールによって一意的に係り先が決まらない場合は, コーパスから得られるデータにより係り先の確率計算をし, その結果により係り先を選択する.

以下, 最初に機械可読辞書から抽出する概念情報を使っての曖昧性解消について述べる. その後で, 選好的曖昧性解消モデル (Preferential Disambiguation Model) を提案し, 選好ルールを述べる. 最後に, この手法によって行った曖昧性解消の実験結果を示し, 本手法の有効性を論ずる.

2　概念情報に基づく曖昧性解消

前置詞句の係り先は文脈に依存する. CIBD では, 曖昧性の解消に用いる文脈を文の中に現れる 4 つの語 (すなわち, 動詞 v, 動詞と前置詞の間の名詞 n1, 前置詞 p, 前置詞の目的語である名詞 n2) に限定する. 以下, この 4 語を (v, n1, p, n2) として参照する.

2.1　曖昧性解消における概念情報の役割

CIBD では, 語の概念分類と語間の概念関係が曖昧性の解消に大きな役割を果す. これらの概念情報を曖昧性の解消に用いるのは, 前置詞句を観察してみると, 次のような一般則が得られるからである.

1. 語の素性がしばしば係り先を決める重要な手がかりとなる. 例えば, n2 が場所で, p が at であれば, 係り先は動詞句である (例：I bought a bottle of wine *at* a *drugstore.*).
2. v と n2, あるいは n1 と n2 の間にある種の概念関係か共起関係があれば, 多くの場合, それにより前置詞句の係り先が決まる. 例えば, Someone had *broken* the window *with* a *stone.* において, *stone*/n2 は *break*/v の道具となるので, この前置詞句は動詞句に係る.
3. n1 と n2 の間に特定の概念関係がある場合, 前置詞句の係り先は名詞句に限定される. 例えば, He spent two years to write a *novel* in three *volumes.* では, *volume* (巻) は *novel* (小説) の構成単位であるため, 前置詞句 in three volumes は名詞 novel に係る.

われわれは v, n1, n2 を特定の意味を持つ概念と考え, それぞれの概念のもつ素性と概念間の意味的関係 (概念関係) を利用して曖昧性を解消する. しかし, この解消法では, 諸種の情報をどこから, どう抽出するかの問題が生ずる. われわれは, EDR 電子辞書を利用してこの問題に対処する. EDR 電子辞書は (株) 日本電子化辞書研究所により編集され, その中に概念辞書, 日本語と英語の単語辞書と共起辞書, 日英と英日対訳辞書, 専門用語辞書, 日本語と英語コーパスを含む (EDR 1993). われわれはその中の概念辞書, 英語の単語辞書とコーパスを使って前置詞の曖昧性を解消する.

2.2　概念類と概念体系

前置詞句の係り先の決定に使うために, 動詞や名詞を素性によりいくつの概念類に分類する. 例えば, 動詞を mental, motion, change_state などの概念類に分類し, 名詞を place, time, state,

abstract, degree, human, animal などの概念類に分類する.

ここで, 概念類は EDR 概念辞書に依拠したものである. EDR 概念辞書は 40 万の概念を持ち, 概念見出し辞書, 概念体系辞書, 概念記述辞書に分けられている. 概念見出し辞書は概念の意味を説明するものである. 概念体系辞書は概念の上位ー下位関係を用いて概念全体をシソーラスにしたものである. 概念記述辞書は概念間の上位ー下位関係以外の意味関係や共起関係による概念を意味ネットワークにしたものである.

一つの概念類は概念辞書の概念体系の中にある一つの概念とその下位概念の集合を意味する. 一つの概念がどの概念類に属するかは EDR 概念辞書の概念体系によって判明する. 図 1 は概念体系の一例を示したものである. ここで, animal という概念類は animal という概念及び animal を上位概念とする下位概念の集合である. dog という概念は animal という概念の下位概念であるので, animal という概念類に所属する (この関係を animal(dog) と書く).

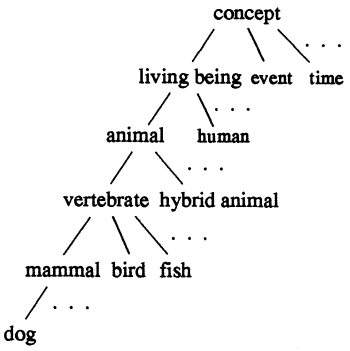

図 1　概念体系の例

2.3　概念関係

v と n2 または n1 と n2 の間の意味関係は前置詞句の係り先の決定に重要な役割を果す. EDR 概念記述辞書は, 概念間の意味関係として 27 種類の概念関係を定義している. 表 1 は, その中から前置詞句の曖昧性の解消に役立つ 12 の概念関係を抽出したものである.

概念記述辞書は二つの概念間に現れる概念関係を記述するものである. 例えば, Tom **repaired** his car with a **wrench**. において, repair(修理する) と wrench(レンチ) の 2 概念の関係は implement となっている. ここで, <repair> − <implement> − ><wrench>は wrench が repair に使う道具であることを示す. したがって, 前置詞句 with a wrench の係り先は repaired である. 同様に, Peter greeted the **girl** in yellow **skirt**. において, girl(若い女) と skirt(スカート) 両概念間 a-object という概念関係がある. これは, skirt は girl がもつものを物語る. 前置詞句 in yellow skirt の係り先はこの関係により名詞 girl に決める.

表 1　曖昧性の解消に用いる概念関係

関係子	記 述	例
a-object	属性を持つ対象	<red>– <a-object> –><apple>
object	動作・変化の影響を受ける対象	<eat>– <object> –><meat>
manner	動作・変化のやり方	<run>– <manner> –><slowly>
implement	有意志動作における道具・手段	<gun>– <implement> –><kill>
source	事象の主体または対象の最初の位置	<come>– <source> –><Osaka>
goal	事象の主体または対象の最後の位置	<go>– <goal> –><Hong Kong>
possessor	所有関係	<father>– <possessor> –><book>
purpose	目的	<go>– <purpose> –><fishing>
condition	事象・事実の条件関係	<delay>– <condition> –><rain>
place	事象の成立する場所	<study>– <place> –><library>
quantity	物・動作・変化の量	<three>– <quantity> –><egg>
number	数	<six>– <number> –><feet>

3　概念情報に基づく選好的曖昧性解消モデル

CIBD では, 概念情報をはじめ, 語彙情報や統語情報や共起情報を用いて曖昧性を解消する. 以下, このモデルの詳細を述べる.

3.1　選好ルール

前置詞句の係り先の決定手順をルール化したものを選好ルール (preference rule) と呼ぶ. 選好ルールは数多くの文例を収集, 分析した結果を一般則にしたものである. 選好ルールはさらに適用範囲によって大域的規則 (global rule) と局所的規則 (local rule) に分けられる. 前者はすべての前置詞に適用され, 後者は特定の前置詞にのみ適用されるものである. 大局的規則は, 構文構造の特徴から抽出した曖昧性解消の手がかりを利用している. 局所的規則は, 概念辞書から抽出した概念情報を利用している.

表 2 に 7 つの大域的規則を, 表 3 に本稿で分析の対象とした 12 の前置詞に対応する局所的規則を示す. 規則の中の $->$ の左側の一項述語は概念類のラベルを示す (例： passivized(v)). 二項述語は概念関係を示す (例： a-object(n1, n2)). これらの規則の右側に, vp_attach は係り先が動詞句, np_attach は係り先が名詞句であることを示す.

表 2　大域的規則

1. lexical(passivized(v) + PP) AND prep ≠ 'by' − > vp_attach(PP)
   v が受身形で前置詞が *by* ではない場合, 前置詞句は VP に係る.
2. n1 == n2 − > vp_attach(n1 + PP)
   n2 と n1 が重複する場合 (例: *step by step,loss on loss*), n1 と PP は一つの構成語になる.
3. (prep ≠ 'of' AND prep ≠ 'for') AND (time(n2) OR period(n2)) − > vp_attach(PP)
   n2 が時を示し, 前置詞が *of* あるいは *for* ではない場合, 前置詞句は VP に係る.
4. lexical(Adjective + PP) − > adjp_attach(PP)
   置詞句の前の語が形容詞 (分詞を含む) の場合, 前置詞句は補語として形容詞に係る.
5. is_a(n2, reflexive_pronoun) − > vp_attach(PP)
   n2 が再帰代名詞の場合, 前置詞句は VP に係る.
6. (v,p) あるいは (v,n1,p) が慣用句として EDR 辞書に登録されている場合, 前置詞句は VP に係る.
7. (n1,p,n2) が慣用句として EDR 辞書に登録されている場合, 前置詞句は NP に係る.

局所的規則を適用する前に, 文の中の単語 v, n1, n2 をそれぞれの意味を表す概念に写像することが必要である. 単語が一つの意味しか持たない場合は, 1 対 1 の写像になる. しかし, 単語は複数の意味を持つ方が通例である. このような場合, 単語を特定の概念に写像することは難しい. ここでは, 1 対 1 の写像をとるのではなく, いくつの意味的に可能な概念に写像する. 以下にそのアルゴリズムを示す.

1. v, n1, n2 の候補概念リストを用意しておく, それぞれの定義を EDR 単語辞書[1]で調べ, とり得る意味 (つまり概念) の集合を各リストに入れる. すべての候補リストに候補概念が一つしかない (単独の意味が特定された) 場合は終了.
2. EDR 英語コーパス[2] に (v, n1, p, n2), (v, p, n2), (n1, p, n2), (v, n1) のどれかと同一の表現が存在するかどうかを調べる. もし存在するなら, 文の中の単語はコーパスの中の単語と同じ意味を持つ可能性が高いため, 候補リストはコーパスの中の概念により置換される. もしすべてのリストに候補概念が一つしかない場合は終了.
3. 一つのリストに二つ以上の候補概念があるとき, 複数の候補が同じ親概念を持つ場合には, それらを親概念によって置換する (これをクラスタリングと呼ぶ).

1 EDR 単語辞書では, 単語や句の意味は特定の概念と対応している.
2 EDR 英語コーパスは 16 万文を含む. 文の形態素情報, 構文情報, 意味情報を提供している.

表 3　局所的規則

about - rules:

motion(v) ⟶ vp_attach(PP)

Default ⟶ np_attach(PP)

at - rules:

construct(n2) OR place(n2) ⟶ vp_attach(PP)

(motion(v) OR transitive(v)) AND object(n2) ⟶ vp_attach(PP)

Default ⟶ np_attach(PP)

by- rules:

implement(v, n2) ⟶ vp_attach(PP)

a_object(n1, n2) OR possessor(n1, n2) ⟶ np_attach(PP)

Default ⟶ np_attach(PP)

for - rules:

goal(v, n2) OR implement(v, n2) OR purpose(v, n2) ⟶ vp_attach(PP)

abstract(n1) ⟶ vp_attach(PP)

Default ⟶ np_attach(PP)

from - rules:

source(v, n2) ⟶ vp_attach(PP)

motion(v) AND (place(n2) OR direction(n2)) ⟶ vp_attach(PP)

Default ⟶ np_attach(PP)

in - rules:

(place(n2) OR construct(n2)) AND !a_object(n1, n2) ⟶ vp_attach(PP)

purpose(v, n2) OR manner(v, n2) ⟶ vp_attach(PP)

movement(v) AND (goal(v, n2) OR place(v, n2)) ⟶ vp_attach(PP)

a_object(n1, n2) OR quantity(n1, n2) ⟶ np_attach(PP)

Default ⟶ np_attach(PP)

of - rules:

change_state(v) and (goal(v, n2) OR object(v, n2)) ⟶ vp_attach(PP)

Default ⟶ np_attach(PP)

on- rules:

condition(v, n2) OR place(v, n2) OR implement(v, n2) OR exchange(v) OR purpose(v, n2) ⟶ vp_attach(PP)

motion(v) AND object(n2) ⟶ vp_attach(PP)

present_participle(n2) ⟶ vp_attach(PP)

place(n1) AND place(n2) ⟶ np_attach(PP)

Default ⟶ np_attach(PP)

over- rules:

implement(v, n2) OR (motion(v) AND (a_object(v, n2) OR goal(v, n2)) ⟶ vp_attach(PP)

Default ⟶ np_attach(PP)

to - rules:

a_object(n1, n2) ⟶ np_attach(PP)

place(n1) AND direction(n2) ⟶ np_attach(PP)

state(n1) AND degree(n2) ⟶ np_attach(PP)

goal(v, n2) ⟶ vp_attach(PP)

Default ⟶ vp_attach(PP)

with - rules:

implement(v, n2) OR manner(v, n2) ⟶ vp_attach(PP)

(a_object(n1, n2) OR possessor(n1, n2)) AND NOT(implement(v, n2) OR manner(v, n2)) ⟶ np_attach(PP)

Default ⟶ vp_attach(PP)

as - rules:

Default ⟶ vp_attach(PP)

3.2 選好的曖昧性解消

CIBD は, 大域的規則と局所的規則を使って前置詞句の係り先を選ぶ. 係り先が一意に決められない場合, 候補となっている係り先の確率計算をし, その値の高いものを係り先として選択する.

CIBD のアルゴリズムは次の通りである.

1. 大域的規則を試す. もし, いずれかの規則が適用可能なら, その規則にある係り先を解として返して終了.
2. 前置詞に関連する局所的規則に現われる単語を前節で述べたアルゴリズムにより概念化する. 未定義語がある場合には, ステップ 4 に行く.
3. 前置詞句に関連する局所的規則を試す. 規則が一つだけ適用される場合は, この規則によって係り先を返して終了. さもなければ, もし前置詞が with でなく, かつ n2 の前に不定冠詞か所有代名詞がある場合, 前置詞句は VP に係る.
4. 次の式で lra-score(Likelihood Ratio on Attachment, 前置詞句が動詞句か名詞句に係る可能性の比率) を算出し, その値により係り先を決める. この値が 1.0 より大きい場合は VP に係る. さもなければ NP に係る.

$$\mathrm{lra(v,p)}=\frac{count(p|vp\_attach)}{count(p|np\_attach)}+log_2\frac{count(v|p)*\Sigma count(prep)}{\Sigma count(v|prep)*count(p)} \qquad (\mathrm{prep} \subset \text{all prepositions})$$

アルゴリズムのステップ 3 で, 係り先が VP か NP の両方になれる場合, n2 の前の修飾語がしばしば係り先を示す. 例えば, 下の例文では, 1a と 2a の前置詞句の係り先は曖昧である. それに対し, 1b と 2b の前置詞句は VP に係ると考えられる.

(1a) Tom cut the meat on the table.
(1b) Tom cut the meat on *a* table.

(2a) They kept the car in the garage.
(2b) They kept the car in *their* garage.

アルゴリズムのステップ 4 の式に用いるデータは EDR 英語コーパスから抽出されたものである. 第一項は指定された前置詞 p における VP と NP に係る割合の比率である. 第二項は動詞と前置詞の相対共起頻度の大きさにより VP に係る傾向性があるかどうかを推定するものである.

lra-score に基づく曖昧性解消はデフォルト手段である. すなわち, 未知語が出てくる場合, あるいは選好ルールにより一意に係り先を決められない場合に統計的手段が使われる.

いくつの例を通して曖昧性解消の過程をみておこう.

(3) I can't find a book suitable for my son.

この文が与えられたとき, まず, 大域的規則を順番に調べる. すると, 四番目の規則が適用されることがわかる. したがって, 前置詞句 *for my son* の係り先は suitable となる.

(4) Suddenly the guest **stopped** her **speech with** a **choking in** her **throat**.

この文には 2 つの前置詞がある. 最初に, 第一の前置詞句 with a choking に対し, 大域的規則の適用を試みる. ここで, どの規則も適用されないことがわかる. 次に, with に関する局所的規則を順番に試みる. 最初の規則により stop/v と choking/n2 の間に implement という概念関係のあることが概念辞書から判明する. また他の規則の適用は不可のため, 係り先は VP に決まる. 第二の前置詞句 in her throat には, 大域的規則は適用できないので, in に関し局所的規則の適用を試みる. ここで三番目の規則が適合する (choking/n1 と throat/n2 の間に scope と a-object という概念関係がある). この規則により前置詞句の係り先は NP となる.

(5) We **extracted** lexical **properties from** a **treebank** to resolve ambiguous PP attachments.

この文には, 大域的規則の適用は不可能である. また, treebank/n2 は辞書に定義されていないので, 局所的規則の適用も不可能である. そこで, lra-score の値を計算すると, 4.733 という値が得られる. この値は 1.0 より大きいので, 前置詞句 from a treebank の係り先は VP となる.

4 実験とその結果

CIBD の有効性を検証するため, 12 の前置詞を含む 2877 の文を使って曖昧性の解消実験を試みた. このテスト用のデータはある新聞と 2 つの本から 4 語組の文をランダムに抽出したものである.

表 4 はその実験結果である. ここでは大域的規則, 局所的規則, lra-score による係り先決定の試行数, それぞれの場合の正解数, 正解率を示す. 大域的規則による係り先決定の正解率は 97.1%に達している. 局所的規則による係り先決定の正解率は 84.4%, lra-score による係り先決定の正解率は 73.9%である. 正解率の平均は 86.5%である.

表 4　CIBD による曖昧性解消の実験結果

段階	試行数	正解数	正解率
大域的規則	784	761	97.1%
局所的規則	1721	1452	84.4%
lra-score	372	275	73.9%
合 計	2877	2488	86.5%

4 語組が EDR 単語辞書で未定義の単語 (未知語) を含む場合, 局所的規則は係り先の決定に適用できない. そこで, lra-score より効率面で優れている局所的規則をできるだけ使えるように、次のようなローカルな処理を施してみることにする.

- 4 桁の数字を年代とみて概念 *date* に置き換える. 他の数字は概念 *number* に置き換える.
- n1 が大文字で始まる文字列 (固有名詞) の場合, それを概念 *name* に置き換える.
- n2 が大文字で始まる文字列の場合, 前置詞が *in* であれば, n2 を概念 *place* に, さもなければ, 概念 *name* に置き換える.
- n2 が人称代名詞の場合, 概念 *person* に置き換える.

これらの処置は Collins らのコーパスの処理と似ている (Collins and Brooks 1995). 表 5 はこの処理を加えた後で同じテストデータを用いて曖昧性解消の実験を試みた結果である.

表 5 未知語の処理を加えた場合の CIBD 実験結果

段 階	試行数	正解数	正解率
大域的規則	784	761	97.1%
局所的規則	1826	1543	84.5%
lra-score	267	197	73.8%
合 計	2877	2501	86.9%

表 5 では, 未知語の処理を加えた正解率が 86.9%に達している. これは表 4 の結果より 0.4%よくなったものである. 言い換えると、この改善は局所的規則で処理不能のケースが 372 文から 267 文に減ったことを意味する.

5 性能の評価

前置詞句の係り先の曖昧性解消実験では, それぞれが用いたテストデータがドメイン, 規模, 処理対象とした前置詞数などの点でを異なっている. したがって, 正解率の精密な比較をすることは難しい. ここでは CIBD が全前置詞を対象にしたときの性能の推定と, CIBD と他の手法との相対的な比較を行っておこう.

5.1 全前置詞を対象としたときの CIBD の性能

CIBD の実験は使用頻度の高い 12 の前置詞を選んで行ったものである. 前置詞を 5 つの独立のテキストで調べてみると, CIBD で対象にした 12 の前置詞は全体の 91.7%(91.4% ~ 92.1%) を占めている. その他の前置詞の出現頻度は 8.3%である. このうちの 27.9%には係り先の決定に

大域的規則を使うことができるので, 表5にある97.1%の正解率を得られよう. 残りの72.1%の前置詞句の係り先をlra-scoreによって決めてみると, 73.8%の正解率が得られると考えられる. このことから, 全前置詞に対しての予想正解率は86.4%となる. [3]

5.2 他の手法との性能の評価

表6は4つの手法によって行った前置詞句の係り先決定の実験結果である. ここで, (1) は構文構造に基づく手法の一つであるRight AssociationをCIBDで用いたテストデータに適用した結果を示す. その正解率は67.1%である. (2) と (3) はコーパスに基づく手法 (Brill and Resnik 1994; Collins and Brooks 1995) の結果である. ここでは, 訓練データとしてThe Wall Street Journal Treebank(Marcus, Santorini and Marcinkiewicz 1993) を使い, テストデータにはIBM Dataを使っている. 正解率はそれぞれ81.9%と84.5%である. (4) は隅田らが提案した用例に基づく手法の実験結果である.

表 6 他の手法とCIBDの比較

手法	データのサイズ	正解率
(1) Right Association	2877	67.1%
(2) Rule-based [BR 94]	3097	81.9%
(3) Backed-off [CB 95]	3097	84.5%
(4) Example-based [SFI 94]	131	85.7%
(5-1) CIBD (付加処理なし)	2877	86.5%
(5-2) CIBD (未知語処理を付加)	2877	86.9%
(5-3) CIBD (全前置詞を対象)		86.4%

ここで, 3299件の人工的に生成した用例 (ドメインは国際会議申込の英日対話文) とLongman Lexicon準拠のシソーラスを用いた曖昧性の解消を討みている. 131件の実験データを使って行った結果の正解率は85.7%である. 本稿での前置詞句の係り先決定結果をこれらの先行実験と比べてみると, CIBDの正解率が他の手法のものより優れていることがわかる. [4]

3 平均正解率 $= 0.917 * 0.869 + 0.083 * (0.279 * 0.971 + 0.721 * 0.738) = 0.864$
4 (2)(3)(5) ともにドメインに依存しないものであり, それらのテストデータの規模には大きな差もない.

6 むすび

CIBDによる曖昧性解消の実験結果は, その有効性を証明した. その理由として次のようなことが考えられる.

1. 機械可読辞書から多様な情報を利用したこと
   CIBDでは, 機械可読辞書と注釈つきコーパスから概念情報をはじめ, 統語情報, 形態素情報, 語彙情報, 共起情報等幅広い情報を入手して, 総合的に曖昧性の解消をしている. 一般に辞書には分野によってカバーする情報の不均一の問題がある. しかし, 各情報の相互補完性を利用することにより, 辞書からの情報が不十分な場合でも, 曖昧性の解消を有効に行っている.

2. 段階的に曖昧性を解消したこと
   CIBDでは, 曖昧性の解消を漸進的に3段階に分け, 成功率の高い方を優先させている. また, 未知語を適切に処理することによって、局所的規則の適用を増やして正解率を上げることに成功している.

3. 多義語への対応をしたこと
   多義語の意味を特定の概念に写像することは難しい. CIBDは, 単語のいくつ意味的可能な概念を選択し, その上位概念へのクラスタリングによって多義語の問題に対処することで正解率を上げている.

本論文は, 概念情報を中心に語彙情報, 統語情報, 共起情報を用いて前置詞句係り先の曖昧性解消手法を示した. CIBDでは, 機械可読辞書とタグつきのコーパスを利用して情報を抽出しているため, 人工的な訓練データを用意する必要がない. また, CIBDはドメインに依存しないため, 汎用性にも優れているといってよいだろう.

参考文献

Boggess, L., Agarwal, R. and Davis, R. (1991). "Disambiguation of Prepositional Phrases in Automatically Labeled technical texts." In *Proceedings of the 9th AAAI*, 155-159.

Brill, E. and Resnik, P. (1994). "A Rule-based Approach to Prepositional Phrase Attachment Disambiguation." In *Proceedings, the 15th COLING*, 1198-1204.

Collins, M. and Brooks, J. (1995). "Prepositional Phrase Attachment Through a Backed-off Model." http://xxx.lanl.gov/cmp-lg/9506021.

Charniak, E. (1993). "Statistical Language Learning." The MIT Press.

Dahlgren, K. and McDowell, J. (1986). "Using Commonsense Knowledge to Disambiguate

Prepositional Phrase Modifiers." In *Proceedings of the 5th AAAI.* 589-593.

Japan Electronic Dictionary Research Institute, Ltd. (1993). "EDR Electronic Dictionary Specifications Guide."

Frazier, L. (1978). "On Comprehending Sentences: Syntactic Parsing Strategies." Doctoral Dissertation, University of Connecticut.

福本文代 (1995). "3 語の同時出現頻度を利用した前置詞句の係り先の曖昧性解消." 自然言語処理, 2(5), 67-74.

Hindle, D. and Rooth, M. (1993). "Structural Ambiguity and Lexical Relations." *Computational Linguistics*, 19(1), 103-120.

Jensen, K. and Binot, J. (1987). "Disambiguating Prepositional Phrase Attachments by Using On-line Dictionary Definition." *Computational Linguistics.* 13(3-4), 251-260.

Kimball, J. (1973). "Seven Principles of Surface Structure Parsing in Natural Language." *Cognition*, 2, 15-47.

Luk, A. K. (1995). "Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions." *Proceedings of the 33rd Annual Meeting of ACL.* 181-188.

Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A. (1993). "Building a Large Annotated Corpus of English: the Penn Treebank." *Computational Linguistics*, 19(2), 313-330.

Nagao M.(1992). "Some Rationales and Methodologies for Example-based Approach." In *Proceedings of FGNLP'92*, 82-94.

Ratnaparkhi, A., Reynar, J. and Roukos, S. (1994). "A Maximum Entropy Model for Prepositional Phrase Attachment." In *Proceedings of the Human Language Technology Workshop*, 250-255.

隅田英一郎, 古瀬 蔵, 飯田 仁 (1994). "英語前置詞句係り先の用例主導あいまい性解消." 電子情報通信学会論文誌 D-II, Vol. J77-D-II, No. 3, 557-565.

Wu, H., Ito, T. and Furugori, T. (1995). "A Preferential Approach for Disambiguating Prepositional Phrase Modifiers." In *Proceedings of the 3rd Natural Language Processing Pacific Rim Symposium*, 745-751.

Whittemore, G., Ferrara, K. and Brunner, H. (1990). "Empirical Study of Predictive Powers of Simple Attachment Schemes for Post-modifiers Prepositional Phrases." In *Proceedings of the 28th Annual Meeting of ACL*, 23-30.

Wilks, Y., Huang, X. and Fass, D. (1985). "Syntax, Preference and Right Attachment." In *Proceedings of the 5th IJCAI*, 779-784.

略歴

呉 浩東: 1983 年重慶大学情報工学部卒業. 1986 年同大学院修士課程修了. 同年, 重慶大学情報工学部助手, 1988 年講師. 1994 年電気通信大学院博士課程入学. 自然言語処理, 知的 CAI, 情報システムの研究に従事.

古郡 廷治: ニューヨーク州立大学計算機科学科博士課程修了. Ph.D. 電気通信大学情報工学科教授. 自然言語処理, 認知科学, 人工知能などの研究に従事. ACM, 情報処理学会, 電子情報通信学会, 計量国語学会各会員.

(1996 年 2 月 13 日 受付)

(1996 年 6 月 14 日 採録)